Letter to the Editor

Potential Structural Motifs for Reverse Transcriptases

Teresa A. Webster, Roberto Patarca,† Richard H. Lathrop,‡ and Temple F. Smith**
*Molecular Biology Computer Research Resource, Dana-Farber Cancer Institute, Department of Biostatistics, Harvard School of Public Health; †Artificial Intelligence Laboratory, Massachussetts Institute of Technology; and ‡Dana-Farber Cancer Institute, Harvard Medical School

Recently, Xiong and Eickbush (1988) generated a detailed primary sequence alignment of RNA-directed DNA polymerases (rd's). The rd's, or reverse transcriptases, are essential enzymes of retroviruses, and recent evidence for reverse transcription has been obtained for other genetic elements, including DNA viruses, transposable elements, and introns (see refs. in Xiong and Eickbush 1988). The Xiong and Eickbush study extends previous work, bringing into a single alignment the rd sequences of these distant elements. Their primary purpose was to reconstruct the potential evolutionary relationship among these sequences. We independently generated a similar alignment for the distantly related viral and LTR-retrotransposon rd's in order to model the potentially conserved structural motif(s). Our approach and result can be viewed as a logical extension of Xiong and Eickbush's work. The combining of sequence similarity with predicted structure to optimize alignments has been reported elsewhere (Nishikawa and Ooi 1986; Webster et al. 1987).

Our alignment of the distant rd's was generated in two steps. First, pairwise sequence comparisons of representative proteins of major clusters were carried out using a modification of the dynamic programming algorithm (Smith and Waterman 1981). This algorithm generates optimal alignments from primary sequence information annotated with neighborhood structural information. Second, the initial sequence alignments, with their predicted secondary-structure annotation, were used to construct complex pattern descriptors for these sequences (for a general discussion of our method for pattern descriptor construction, see Webster et al. 1987, 1988). All descriptor matches within the rd sequences and to the entire NBRF/PIR (George et al. 1986) version 16 data base were identified. Finally, for each descriptor, the sensitivities (% rd's that match the descriptor) and specificities (% non-rd's in the NBRF/PIR version 16 data base that do not match the descriptor) were estimated.

On the basis of this analysis, there appear to be at least four common blocks of structural similarity among distant rd sequences (fig. 1). These blocks are labeled 2–5 (following the nomenclature of Xiong and Eickbush). They are displayed relative to the sequence of HIV-1 (Sanchez-Pescador et al. 1985) in figure 1. Block 5 is the most strongly conserved and has elsewhere been characterized as a DD dipeptide followed by several hydrophobic residues (reviewed in Baltimore 1985). Our optimized pattern for seven amino acids of this region of block 5 is [WYF]-[ILVMWYFC]DD[ILV][ILVMWYFC][ILVMWYFC], with predicted beta strands flanking the DD dipeptide within four amino acids (see legend to fig. 1). This descriptor has an estimated 100% sensitivity and 100% specificity for the rd's. Given the shortness of the loop, this region is likely to form a beta hairpin (Milner-White and Poet 1987; Argos 1988).

A second pattern descriptor (see legend to fig. 1) characterizes blocks 2–4 with a sensitivity of 80% and a specificity of 99.9%. Inclusion of the secondary-structure elements in the descriptor increases its sensitivity and specificity correlation coefficient by 23%. Alignment of the regions of the rd's (listed in the legend to fig. 1) that match the two

Address for correspondence and reprints: Dr. Teresa Webster, Molecular Biology Computer Research Resource, Dana-Farber Cancer Institute, Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115.

Mol. Biol. Evol. 6(3):317–320. 1989.

0737-4038/89/0603-008$02.00

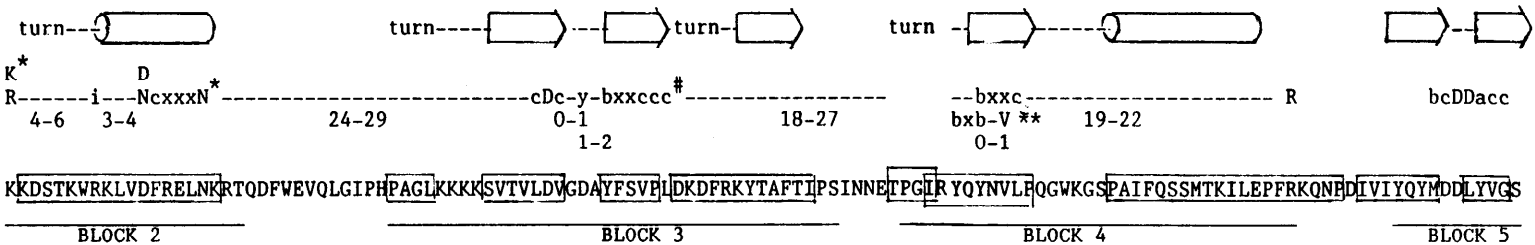

FIG. 1.—Potential structural motifs for viral and LTR-retrotransposon rd's. The elements of the complex pattern descriptors are shown in the top two lines. The top line shows the sequence of predicted secondary-structure elements, where the arrow denotes beta strand and the cylinder denotes alpha helix. The second line from the top shows the sequence of primary elements, where i − j (written below this line) = the allowed ranges of amino acids of any type between adjacent primary elements. Symbols are as follows: y = [HRKSTQNED]; a = [ILV]; b = [YWF]; c = [MILVPCWYF]; i = [HKR], where the brackets enclose a group of allowed amino acids; uppercase letters = one-letter amino acid code; x = any amino acid. One representative rd sequence, that of HIV-1 (Sanchez-Pescalor et al. 1985; PIR locus = GNVWA2), is aligned with the descriptor's primary and secondary elements. Amino acid regions that are predicted to fold into the corresponding secondary-structure elements are enclosed in boxes. Searches for matches to the block 2–4 descriptor were carried out by identifying regions that match the primary element pattern, extracting these regions plus 20 flanking amino acids, annotating the extracted regions with predicted secondary-structure elements, and then searching for matches to the following pattern: [KR] $X_{0,1}$ turn $X_{0,3}$ $X_{0,16}$ turn $X_{-2,6}$ beta $X_{-4,0}$ cDc $X_{-5,10}$ y $X_{1,2}$ bc $X_{-3,-2}$ beta $X_{-2,5}$ turn $X_{-1,4}$ beta $X_{0,11}$ turn $X_{-1,3}$ bXXc $X_{-8,-3}$ beta $X_{-2,17}$ alpha, where beta = beta strand; alpha = alpha helix; and turn = beta turn and where the $X_{i,j}$ (spacer) is an allowed range of amino acids of any type between flanking elements with a negative value representing an allowed overlap between flanking elements. For example, in the pattern beta $X_{-6,2}$ alpha, the N terminus of the alpha element can be located between the position that is five amino acids to the left—and the position that is three amino acids to the right—of the position occupied by the C terminus of the beta element. The pattern descriptor constructed for block five is beta $X_{-2,0}$ [YFW][MILVYWFC]-DD[LIV][MILVYWFC][MILVYWFC] $X_{-3,1}$ beta. The pattern descriptor constructed for ssRNA binding is as follows: turn $X_{-2,4}$ bxb $X_{-7,0}$ beta. The secondary-structure elements were predicted by a modification of the Chou/Fasman algorithm (Ralph et al. 1987) using default parameters, with the exception of the betas in block five for which beta threshold = 1.21. Searches for matches to complex pattern descriptors were carried out by the program ARIADNE (Lathrop et al. 1987). All mismatched secondary-structure elements were given a score of −1. Descriptor sensitivities were estimated using the following 10 groups of rd sequences: DIRS1; 412, GYPSY,17.6; CAMV; HIV1, HIV2, VISNA, EIAV; SRV, IAP,RSV; BLV, HTLV1, HTLV2; MuLV; HBV, DHBV; Copia; and Ty (for references and nomenclature, see Xiong and Eickbush 1988). The asterisk (*) denotes that either the [RK] or the N alone is sufficient to satisfy the descriptor; the number sign (#) denotes that two of the three must be "c" elements; the double asterisks (**) indicate the primary sequence of descriptor for ssRNA binding.

descriptors results in a multiple sequence alignment that is essentially identical to the one generated independently by Xiong and Eickbush. The only exceptions are the two most distant rd's, those of copia (mobile element from *Drosophila melanogaster*) and Ty (mobile element of *Saccharomyces cerevisiae*). For these sequences we bring into the alignment the regions that match the primary elements of blocks 2 and 4. As in the Xiong and Eickbush study, we find that rd's of copia and Ty are the most divergent throughout this domain, while the rd's of HBV (human hepatitis B virus) have no match to block 2 and require a very large insertion to align with block 3. All regions that match the intrablock connective regions of descriptor 2 (fig. 1) are predicted, by the PLANS algorithm, to be at the surface of the tertiary structure of the protein (Cohen et al. 1986). The spacing of predicted intervening surface regions suggests that all four conserved blocks can be folded together to form a complex active site.

We asked, What subfunctions could possibly correlate with these conserved structural blocks, given the minimum required rd functions: nonspecific binding to an RNA template, binding to all four deoxynucleotides, the stabilization of a phosphodiester bond in the transition-state complex, and perhaps binding to an RNA-DNA heteroduplex? Placement of block 2 as being at or near the deoxynucleotide binding site is suggested by a study in which the first basic amino acid in this block for the MuLV rd was cross-linked with an agent specific for the triphosphate binding site (Basu et al. 1988). The supposition that the strongly conserved and essential (Larder et al. 1987) block 5 is involved with one or more of these subfunctions is supported by the fact that a similar sequence exists in a very different class of polymerases—the RNA-directed RNA polymerases (rr's) (Kramer and Argos 1984; Argos 1988). As with the rd's, we predict that this region in the rr's folds into a beta hairpin, with the DD dipeptide forming the loop. A single descriptor for this potentially common beta hairpin characterizes all RNA-directed polymerases with 92% sensitivity and 99% specificity. The common subfunctions across these two polymerase classes are the nonspecific binding to an RNA template and the formation of the phosphodiester bond.

Since the elements of blocks 2–5 are conserved among highly divergent rd sequences, we examined whether they may in fact form components of other polymerase active sites. However, the potentially common beta hairpin regions of the rr's and rd's appear to contain the only similarities linking the two RNA-directed polymerase classes (authors' personal observation). In addition, we find no matches to our rd pattern descriptors within two other polymerase classes, the DNA-directed DNA polymerases and the DNA-directed RNA polymerases. However, a variant of the above beta hairpin—a variant in which the loop is formed by the sequence DTD rather than by DD—may exist within the DNA-directed DNA polymerases (Argos 1988). Our analysis rules out a previously noted similarity (Johnson et al. 1986) between the block 5 region of rd's and the alpha subunit of *Escherichia coli* DNA-directed RNA polymerase. We find that this similarity is not conserved among homologous alpha-subunit sequences from tobacco and liverwort chloroplasts and does not contain predicted beta strands that closely flank the DD dipeptide.

Finally, a potential function for block 4 as part of the RNA template binding site is suggested by our discovery of a similarity between a subcluster of the rd's—which includes the rd of CaMV and all retroviral rd sequences except MuLV—and five sequences of two RNA binding proteins, the A1 RNP core protein and the poly(A) binding protein. These five regions were elsewhere noted to be putative, nonspecific, ssRNA binding sites (Chung and Wooley 1986). The similarity consists of the following sequence: [aromatic] X [aromatic] $X_{0,1}$ valine (where "X" is any amino acid), contained within a beta strand and proceeded by a beta turn (fig. 1). A pattern descriptor (see legend to fig. 1), which matches both the subcluster of rd's and the repeats within the ssRNA binding proteins, has 97% specificity against the nonpolymerase sequences in the PIR version 16 data base. These results and the observation that aromatic groups intercalate within the nucleic acid chain suggest that this motif in the rd's

interacts with the RNA template. Among the nonretroviral rt's, divergence from the descriptor includes the loss of both the first aromatic group and the valine. However, all nonviral rd's retain the middle aromatic group as well as the pattern of predicted secondary-structure elements. Thus, this central aromatic group may make an essential contact with the RNA, perhaps via a ring-stacking interaction.

Note added in proof: We have identified 85% of the 34 elements of the eukaryotic-related structural motif in the newly sequenced prokaryotic reverse transcriptase of Lampson et al. (1989).

LITERATURE CITED

ARGOS, P. 1988. A sequence motif in many polymerases. Nucleic Acids Res. **16**:9909–9916.

BALTIMORE, D. 1985. Retroviruses and retrotransposons: the role of reverse transcription in shaping the eukaryotic genome. Cell **40**:481–482.

BASU, A., V. B. NANDURI, G. F. GERARD, and M. J. MODAK. 1988. Substrate binding domain of murine leukemia virus reverse transcriptase. J. Biol. Chem. **263**:1648–1653.

CHUNG, S.-Y., and J. WOOLEY. 1986. Set of novel, conserved proteins fold pre-messenger RNA into ribonucleosomes. Proteins **1**:195–210.

COHEN, F. E., R. M., ABARBANEL, I. D., KUNTZ, and R. J. FLETTERICK. 1986. Turn prediction in proteins: a complex pattern matching approach. Biochemistry **25**:266–275.

GEORGE, D. G., W. C. BARKER, and L. T. HUNT. 1986. The protein identification resource (PIR). Nucleic Acids Res. **14**:11–15.

JOHNSON, M. S., M. A. MCCLURE, D.-F. FENG, J. GRAY, and R. F. DOOLITTLE. 1986. Computer analysis of retroviral *pol* genes: assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes. Proc. Natl. Acad. Sci. USA **83**:7648–7652.

KRAMER, G., and P. ARGOS. 1984. Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. Nucleic Acids Res. **12**:7269–7282.

LAMPSON, B. C., J. SUN, M. HSU, J. VALLEJOR-RAMIREZ, S. INOUYE, and M. INOUYE. 1989. Reverse transcriptase in a clinical strain of *E. coli:* production of branched RNA-linked msDNA. Science **24**:1033–1038.

LARDER, B. A., D. J. M. PURIFOY, K. L. POWELL, and G. DARBY. 1987. Site-specific mutagenesis of AIDS virus reverse transcriptase. Nature **327**:716–717.

LATHROP, R. H., T. A. WEBSTER, and T. F. SMITH. 1987. ARIADNE: pattern-directed inference and hierarchical abstraction in protein structure recognition. Commun. Assoc. Comput. Machinery **30**:909–921.

MILNER-WHITE, E. J., and R. P. POET. 1987. Loops, bulges, turns and hairpins in proteins. Trends Biochem. Sci. **12**: 189–192.

NISHIKAWA, K., and T. OOI. 1986. Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. Biochem. Biophys. Acta **871**:45–54.

RALPH, W. W., T. WEBSTER, and T. F. SMITH. 1987. A modified Chou and Fasman protein structure algorithm. CABIOS **3**:211–216.

SANCHEZ-PESCADOR, R., M. D. POWER, P. J. BARR, K. S. STEIMER, M. M. STEMPIEN, S. L. BROWN-SHIMER, W. W. GEE, A. RENARD, A. RANDOLPH, J. A. LEVY, D. DINA, and P. A. LUCIE. 1985. Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). Science **227**:484–492.

SMITH, T. F., and M. S. WATERMAN. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

WEBSTER, T. A., R. H. LATHROP, and T. F. SMITH. 1988. Pattern descriptors and the unidentified reading frame 6 human mtDNA dinucleotide-binding site. Proteins **3**:97–101.

WEBSTER, T. A., R. H. LATHROP, and T. F. SMITH. 1987. Prediction of a common structural domain in aminoacyl-tRNA synthetases through use of a new pattern-directed inference system. Biochemistry **26**:6950–6957.

XIONG, Y., and T. H. EICKBUSH. 1988. Similarity of reverse transcriptase–like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol. **5**:675–690.

WALTER M. FITCH, reviewing editor

Received December 5, 1988; revision received December 12, 1988